Implementation of complex interactions in a Cox regression framework

Martina Müller[1] and Kurt Ulm

Institute for Medical Statistics and Epidemiology, IMSE
Technical University Munich, Ismaningerstr.22, 81675 München, Germany

Abstract

The standard Cox proportional hazards model has been extended by functionally describable interaction terms. The first of which are related to neural networks by adopting the idea of transforming sums of weighted covariables by means of a logistic function. A class of reasonable weight combinations within the logistic transformation is described. Apart from the standard covariable product interaction, a product of logistically transformed covariables has also been included in the analysis of performance of the new terms. An algorithm combining likelihood ratio tests and AIC criterion has been defined for model choice. The critical values of the likelihood ratio test statistics had to be corrected in order to guarantee a maximum type I error of 5% for each interaction term. The new class of interaction terms allows interpretation of functional relationships between covariables with more flexibility and can easily be implemented in standard software packages.

Keywords: *Survival analysis, Cox proportional hazards, modelling of interactions*

[1] martina.mueller@imse.med.tu-muenchen.de

1 Motivation

One of the main interests of statistical medical research is the detection of prognostic factors and judgement of their impact on well known diseases. The Cox proportional hazards model (Cox, 1972) has become a standard method for analysing multivariate survival data.
Interactions between covariables are commonly introduced as a product of the two variables of interest which obviously sometimes is naive because they can potentially be much more complex. Therefore, detection and more detailed description of interactions has become a focus of scientific research. Different methods have been used to describe interactions. In a survival context, interactions between covariables and observation time are of special interest as, if they occur, the assumption of proportional hazards for application of Cox's regression method is violated. Among the methods applied in this context are varying coefficient models, introduced by Hastie and Tibshirani in 1993. Their approach allows modelling the influence of one variable in dependence of another variable. This method can be applied to model time dependencies as well as interactions between the covariates. Another method, presented by Grambsch and Therneau in 1995, is based on smoothed scaled Schoenfeld residuals. Some further methods described are: application of fractional polynomials (Berger et al., 2002), isotonic regression algorithms (Salanti & Ulm, 2003) and a tree-based approach (Xu & Adak, 2002).
One must consider the form of the covariates in the interaction term. For an interaction between a continuous covariate and a factor the easiest way is to compute the covariates influence separately for all the factor levels by introduction of dummy variables. An example for a test on the interaction between treatment and continuous covariates based on fractional polynomials has been presented by Royston in 2002.
Calculation of the interaction surface for two continuous covariates is more complicated. As mentioned above, varying coefficients (Hastie and Tibshirani, 1993) can be employed. Lang and Brezger (2003) present a method to approximate the surface by a tensor product of bayesian p-splines. In 1999, LeBlanc and Crowley developed a survival version of multivariate adaptive regression splines (MARS) using weighted least squares and draw a connection to the earlier described method of smoothed martingale residuals (Therneau et al., 1990) whereas the use of the latter is not recommended in case of correlation (Therneau & Grambsch, 2000) and therefore for interaction terms in presence of their corresponding main effects. Locally constant surfaces can be created using CART (classification and regression trees) algorithms (Zhang & Singer, 1999) or isotonic regression (Salanti, 2003).
Recently, there have been several attempts to introduce neural networks to medical statistics. These achieve high flexibility by introducing so called *hidden layers* consisting of *hidden units* where the input units, e.g. the covariates, are summed, transformed and passed to the next layer until the

output layer is reached. This approach allows for high dimensional interactions. *Skip layer connections* are also allowed. In this way, simple linear terms can be introduced. A neural network with one hidden layer and three hidden units can appear as below:

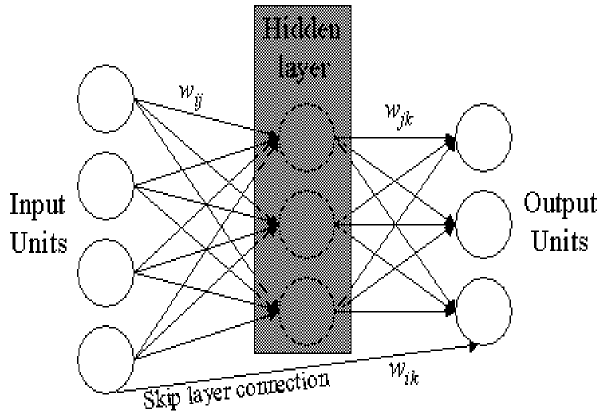

Figure 1. An example of a neural network with five input units, i.e. covariables, one hidden layer consisting of three hidden units, three output units and one skip layer connection. The signal is passed along the arrows and multiplied by the corresponding weight w. The sum of incoming signals at each unit is usually transformed logistically. This transformation function is called activation function and can have various shapes although, in most cases, the logistic function is preferred.

The formula for output unit k of a neural network with one hidden layer and skip layer connections writes:

$$y_k = \phi_{out} \left(\sum_{i \to k} w_{ik} x_i + \sum_{j \to k} w_{jk} \phi_h \left(\sum_{i \to j} w_{ij} x_i \right) \right) \quad (1)$$

The index i stands here for the input units, j is for the hidden layer units and k is for output units. The transformation functions ϕ are commonly chosen as logistic functions. The weights w must be optimised. Details about statistical neural networks and their optimisation can be found for example in Anders (1997) or Ripley (1996).

Ripley (1998) compared applications of neural networks for different survival models. Unfortunately, the high flexibility which is achieved in modelling leads to problems interpreting the actual effects of the original input units, the covariates, and describing the functional form of the interaction terms (Schwarzer et al., 2000). Additionally, neural networks can not take censored time data as output. Therefore, various approaches have been presented to circumvent this problem, e.g. survival times are estimated using a standard

Cox model and fed into the neural network or the output variable is split into several time intervals. This leads to another major problem of neural networks. There are weights on all connections between input, hidden and output units which must be optimised. There is a high danger of overfitting, this is why they can only be used efficiently for large data sets. Medical data is expensive and therefore most data sets are quite small.
Although a lot of work has been done to describe interaction surfaces, there is no general way to approximate interaction surfaces functionally. Consequently, standard medical research still relies on product interaction terms. The understanding of the functional form or at least a good approximation, could help to understand the underlying mechanisms of diseases. As already mentioned above, medical data is difficult and expensive to acquire, therefore, data sets for analysis are generally small. If a functionally describable interaction can be found once for a disease it would be easy to check for the same relationship in different data sets for the same disease avoiding computer intensive methods and specialist knowledge.
The aim of this work is to find an alternative to the standard product interaction term within a Cox regression framework which is based, more or less, on complex functional forms and still allows for interpretation.

2 Methods

2.1 Survival analysis

Often the outcome for a medical analysis is the time to an event such as death or relapse. However, the event does not necessarily occur during the time that the patient is under observation. These patients are censored at the end of the study. Furthermore, some patients drop out during the study for various reasons, such as: refusal to continue, or death for reasons other than their disease. In the latter case, some researchers argue that the underlying reason for death, for example for an accident, is still strongly related to the disease. Therefore these people are sometimes still treated as subjects with event. In general however, the dropouts are also censored observations.
The Cox proportional hazards regression model introduced by Cox (1972) gives a solution for dealing with censored time data. The hazard function $h(t)$ is defined as the instantaneous probability of an event at time t given the individual has survived until time t. The Cox proportional hazards model is defined as:

$$h(t) = h_0(t) \exp(X\beta) \tag{2}$$

No assumptions are made for the shape of the underlying hazard $h_0(t)$. A difference in X only results in a constant shift of the underlying hazard

function. For each of the covariates X_i in X the factor $\exp(\beta_i)$ gives a measure called *relative risk* and can be interpreted as the shift in the hazard function as result of a shift of one unit in the corresponding covariable X_i. This proportionality assumption is the reason for the model's name: proportional hazards model.
Estimation of parameters β within a Cox regression framework with untied failure times is based on the partial likelihood function introduced by Cox (1972) which is the first factor of the full likelihood and is not dependent on the underlying baseline hazard $h_0(t)$. Cox also showed that this partial likelihood is appropriate for estimation. It is defined as follows:

$$L(\beta) = \prod_{Y_i uncensored} \frac{\exp(X_i\beta)}{\sum_{Y_j \geq Y_i} \exp(X_i\beta)} \tag{3}$$

As can be seen, this partial likelihood function only depends on the parameters β. In the case of tied failure times a correction for the partial likelihood would be required. Breslow (1974) proposed the following correction:

$$L(\beta) = \prod_{Y_i uncensored} \frac{\exp(s_n\beta)}{\sum_{Y_j \geq Y_i} \left[\exp(X_i\beta)^{d_n}\right]} \tag{4}$$

In this formula, d_n is the number of failures at failure time t_n and s_n is the sum of corresponding individuals' covariables. In the presence of too many tied failure times a discrete model is preferable (Fahrmeir et al., 1996).

2.2 Interaction structure

The first idea for interaction modelling is loosely based on the structure of neural networks. Each connection between the different units of a neural network can be thought of as neuron which fires e.g. passes information to the next layer or does not. This is usually realised by a logistic transformation of the summed weighted input signals.

$$x_{next.layer} = f_{logistic}\left(\sum_i weight_i * x_{prev.layer.unit_i}\right) \tag{5}$$

where

$$f_{logistic}(x) = \frac{1}{1+\exp(-x)} \tag{6}$$

This transformation function is called activation function and need not be specifically a logistic function. There are many different shapes which have been used in statistical literature that begin with simple step functions (Duch & Jankowski, 1997).
The idea of a smooth threshold function for interaction terms is realistic in a medical context as a change of risk is often monotonic. Consequently,

continuous prognostic factors are often recoded as binary variables, dividing the patients into a low risk and a high risk group. Interpreting this idea for interaction terms would mean that combined effects only arise when the two interacting variables exceed their associated threshold values. In contrast to underlying step functions the resulting interaction surface would be smooth. Steepness and position of the steps should be flexible within the model choice procedure.
Hence, the two interacting variables, as in neural networks, are weighted, summed, and logistically transformed. The weights are chosen as the best fit from a carefully chosen predefined set of weights. The new complex interaction term is written as:

$$f_c\left(x_1, x_2\right) = f_{logistic}\left(w_1 x_1 + w_2 x_2\right) \tag{7}$$

Another possibility which is closer to the standard procedure of the covariable product and still provides a sloping surface is to transform the covariables logistically before they are multiplied. The resulting interaction term is indexed $t.m$ which means *transformed - multiplied*:

$$f_{t.m}\left(x_1, x_2\right) = f_{logistic}\left(x_1\right) * f_{logistic}\left(x_2\right) \tag{8}$$

The last interaction term discussed herein is the standard covariable product:

$$f_m\left(x_1, x_2\right) = x_1 * x_2 \tag{9}$$

The resulting surface has the form of a valley if the included covariates have positive as well as negative values (see figure 3).
The three interaction surfaces are treated as competing models from which the best is chosen.

2.3 Weight choice

First, two standard normally distributed variables were simulated to choose a reasonable set of weights within the new interaction transformation. Standardised covariables were chosen to guarantee smoothness of the surface. The logistic function transforms all variables with an absolute value of more than 2.95 to values between [0;0.05] for negative values and [0.95;1] for positive values. This is visualised in figure 2, where the logistic function is shown together with a standard normal distribution density plot.

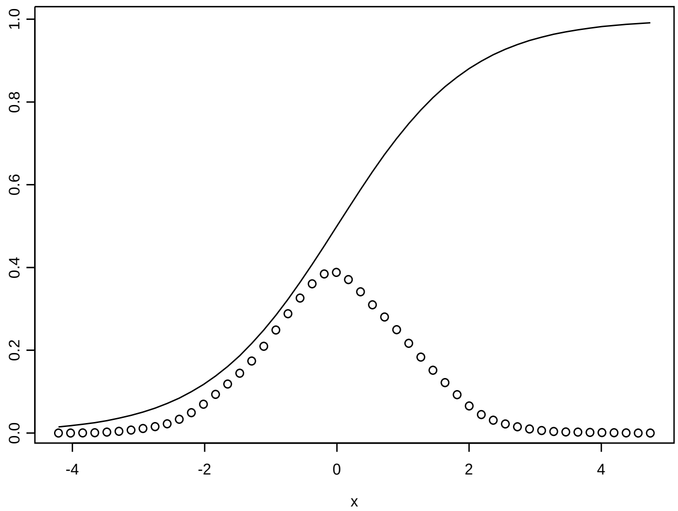

Figure 2. Logistic function and standard normal distribution density. The solid line shows logistically transformed values of x. The circles are density values of a N(0;1) distributed variable. The units on the y-axis are therefore interpreted as values of a logistic function in one case and as density values in the other. If x is standard normally distributed the values resulting from a logistic transform will range between 0 and 1 without clustering at either of the ends.

After plotting the surfaces, the weights in question should be among combinations of $\{-2, -1, -0.5, 0.5, 1, 2\}$. Further increase of the absolute values leads to similar interaction surface shapes to those created using an absolute value of 2. Smaller distances between the weights can hardly be detected. Furthermore, there are redundant combinations as shown by:

$$f_{logistic}(x) = 1 - f_{logistic}(-x) \tag{10}$$

Consequently, negatives of chosen combinations only affect the offset, which means in survival context the baseline hazard, and should be excluded as well as multiples. For example (2;2) is excluded if (1;1) already is in the set.

Further investigation showed that combinations between 2 and 1 resulted in similar interaction surfaces as combinations of 2 and 0.5. Therefore the finally chosen set of weight combinations is

$$\{(2; 0.5), (0.5; 2), (2; -0.5), (-0.5; 2), (1; 1), (1; -1)\} \tag{11}$$

The resulting interaction surfaces for all allowed weight combinations and the alternative models are shown in figure 3. Although some of the surfaces look quite similar they can be distinguished from each other as shown in Chapter 2.5.

2.4 Critical values

Data simulation

In a simulation study 20000 survival data sets were generated each of which consisted of 1000 observations. The covariables are standard normally distributed which refers to standardised covariables. The linear predictor is chosen as:

$$lin.pred. = x_1 - 0.5 * x_2 \tag{12}$$

The relative risk is defined as:

$$rr = exp(lin.pred.) \tag{13}$$

For each observation an exponentially distributed survival time was simulated with expected value $1/rr$. As the resulting survival times are very small, they were multiplied by 365 changing the time unit from years to days. Additionally a uniformly distributed censoring time was simulated for the interval $[0; 1000]$. This simulates a random censoring process and a maximum time of observation of 1000 days. For each observation, the minimum of the two simulated times was taken as observation time, and the event indicator was set to 1 for each survival that is shorter than the censoring time, otherwise the indicator is set to 0.
For the 20000 simulated data sets the mean censoring percentage is 37.83% with standard deviation 0.015. On each of the generated 20000 data sets the four different Cox proportional hazard models where tested. The first model was the true model without any interaction term. The following models contained one of the earlier described interaction terms:

- $f_m(x_1, x_2) = x_1 * x_2$
- $f_{t.m}(x_1, x_2) = f_{f_{logistic}}(x_1) * f_{f_{logistic}}(x_2)$
- $f_c(x_1, x_2) = f_{f_{logistic}}(w_1 x_1 + w_2 x_2)$

The hazard function for a Cox model in our context has the following form:

$$h(t) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \beta_{int} f(x_1, x_2)) \tag{14}$$

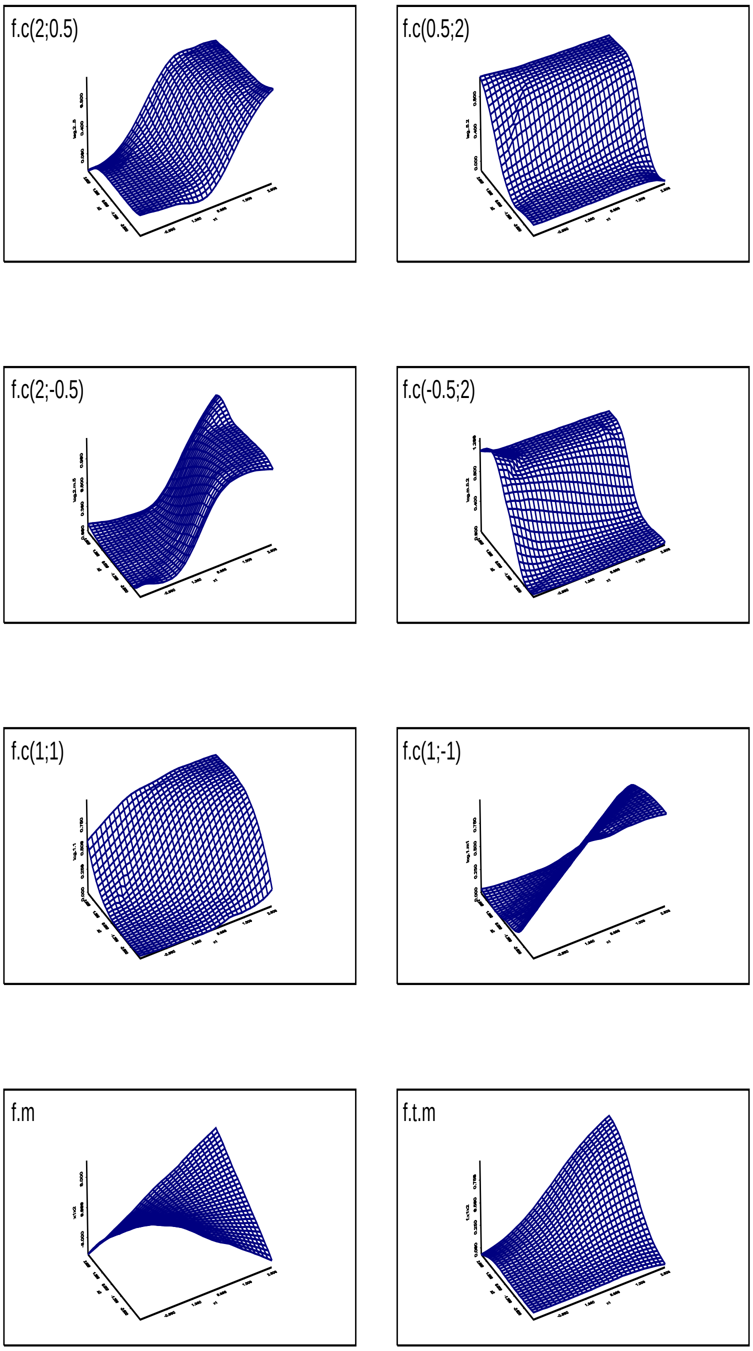

Figure 3. Interaction surfaces considered in the study generated based on two standard normally distributed variables. The upper six plots show surfaces for the newly introduced complex interaction term $f_c(x_1, x_2) = f_{logistic}(w_1 x_1 + w_2 x_2)$. The weights w_1 and w_2 are indicated in the associated figure titles. The lower two surfaces show the standard variable product's surface ($f_m(x_1, x_2) = x_1 * x_2$) and the surface resulting from a product of logistically transformed variables $f_{t.m}(x_1, x_2) = f_{logistic}(x_1) * f_{logistic}(x_2)$.

Another subprocedure was introduced to choose the best weight combination within the complex interaction model. The six possible submodels do not differ in their number of degrees of freedom. Thus, the combination which yielded the highest likelihood was chosen.
In the next step, for each interaction term the optimal critical value of the likelihood ratio test statistic was computed in order to guarantee the desired significance level of 95%. This is achieved by calculating the 95% quantile of the 20000 likelihood ratio test statistics for inclusion of the corresponding interaction term in the model.
In case of more than one significant interaction model the final choice was based on the AIC criterion as the models are not nested and differ in the number of degrees of freedom. The AIC criterion is defined as:

$$AIC = -2 * loglik + 2 * DF \tag{15}$$

The complex interaction model needs four degrees of freedom as the weights are chosen from a predefined set of combinations whereas the other interaction models only need three and the main effect model needs only two. The lowest AIC was the indicator for the best model.

Results

The critical values for 5% error rate of each tested interaction term does not equal the expected X^2 distribution values.
For the product term as well as the product of the transformed variables one would expect a χ^2 distribution with one degree of freedom. The corresponding number of wrong decisions in the generated data would be 8.075% for the multiplicative and 8.25% for the transformed multiplicative term. For the new complex interaction term, $f_c(x_1, x_2)$, a χ^2 distribution with two degrees of freedom would lead to 8.005% wrong decisions. Consequently the critical values c_m, $c_{t.m}$ and c_m must be adjusted.
Dividing the 20000 calculated likelihood ratio test statistics for each of the interaction models into 20 arbitrary groups of 1000 and calculating 20 different 95% quantiles gave a measure for deviation and therefore a 95% confidence interval for the obtained values as shown in table 1. The mean of the 20 critical values differed negligibly from the 95% quantiles of all the 20000 values. The expected χ^2 values were not even in the calculated 95% confidence region.

Table 1. New critical values for interaction terms

Interaction term	**Critical value**	**95% CI**	χ^2 **value for 95%**
$f_m(x_1, x_2)$	4.714338	(4.485082; 4.898646)	3.841459 (1 DF)
$f_{t.m}(x_1, x_2)$	4.792402	(4.54064; 5.141197)	3.841459 (1 DF)
$f_c(x_1, x_2)$	6.918027	(6.689027; 7.141436)	5.991465 (2 DF)

The new critical values were also tested for different censoring percentages. This was realised by varying the upper limit of the uniformly distributed censoring time which was originally set to 1000. The new censoring limits (opposed to 1000) were: {100; 200; 300; 500; 2000; 5000; 10000 }. For each of which another 20000 data sets were simulated. The resulting mean censoring percentages and the corresponding 95% quantiles of the likelihood ratio test statistics for the different interaction terms are shown in table 2 and visualised in figure 4.

Table 2. 95% quantiles of likelihood ratio test statistic for different censoring percentages

cens. limit	**100**	**200**	**300**	**500**	***1000***	**2000**	**5000**	**10000**
% cens.	82.2%	71.4%	63.6%	52.9%	*37.8%*	24.5%	12.0%	6.5%
f_m	3.926691	4.008853	4.132055	4.405948	*4.714338*	4.659497	4.507719	4.484139
$f_{t.m}$	3.876166	3.996379	4.024657	4.454415	*4.795741*	4.6715	4.499799	4.47687
f_c	6.77715	6.654635	6.741761	6.813192	*6.918027*	6.769413	6.610614	6.645809

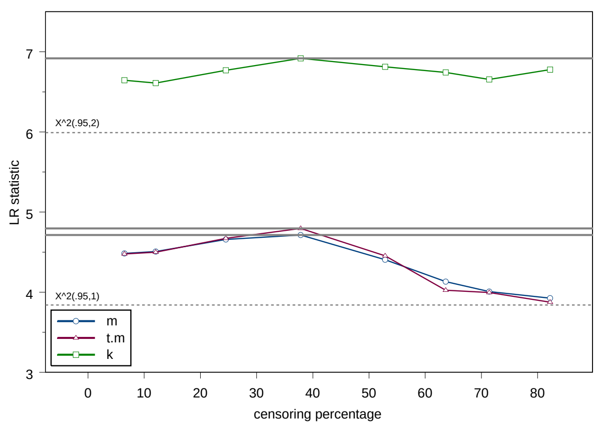

Figure 4. Limits of likelihood ratio test statistic for different censoring percentages. Expected χ^2 values are indicated as dashed reference lines, maximum values as solid lines

The solid lines in figure 4 represent the new critical values chosen in first place for 37.83% censored observations. They are the strictest while only for more than 80% censoring the two multiplicative models' critical values, which behave quite similar to each other, approach the standard χ^2 value - represented here as a dashed line.
For the new term the χ^2 value for two degrees of freedom and 95% probability would lead to minimum 6.87% type I error. The calculated critical value does not approach the expected χ^2 value for any of the censoring rates considered.
Therefore, to guarantee a maximum type I error rate of 5% within the simulated data, the critical values referring to the data with 37.83% censoring were used for the following studies. The corresponding errors for the other censoring rates were calculated and are displayed in table 3.

Table 3. Type I error rates corresponding to the new critical values for different censoring rates

% censored	82.2 %	71.4%	63.6%	52.9%	*37.8%*	24.5%	12.0%	6.5%
Error f_m	3.19%	3.4%	3.58%	4.205%	*5%*	4.85%	4.46%	4.35%
Error $f_{t.m}$	3.06%	3.225%	3.18%	4.02%	*5%*	4.675%	4.275%	4.195%
Error f_c	4.64%	4.315%	4.525%	4.775%	*5%*	4.725%	4.23%	4.31%

Note that these values are only for some censoring percentages. For further studies, the exact relationship between the critical values and the censoring percentage should be determined in order to use the correct values or at least to find the maximum critical value.
When the likelihood ratio test and the AIC criterion indicate that the complex interaction model is the best model then the weight combination determined during the procedure is examined in detail (table 4). Note that the percentages in table 4 are rounded. Therefore, the line sums do not add to 100% as indicated in the last column. The percentages in table 4 are calculated within the number of chosen complex interaction models which make up no more than 5% of the simulated data per censoring percentage, which is no more than 1000 models.

Table 4. Weight combinations for chosen complex interaction terms in case of no underlying interaction

% censoring	$(0.5; 2)$	$(2; 0.5)$	$(1; 1)$	$(1; -1)$	$(-0.5; 2)$	$(2; -0.5)$	$\sum$
82.2%	16.71%	17.52%	20.68%	13.55%	16.94%	14.60%	100%
71.4%	16.12%	17.00%	14.99%	15.37%	18.51%	18.01%	100%
63.6%	17.52%	19.19%	15.85%	15.61%	16.57%	15.26%	100%
52.9%	16.91%	17.02%	18.49%	17.47%	15.45%	14.66%	100%
37.8%	16.22%	17.08%	13.96%	19.55%	16.43%	16.76%	100%
24.5%	17.18%	16.95%	17.63%	17.18%	14.58%	16.50%	100%
12.0%	15.63%	17.13%	15.00%	19.75%	17.25%	15.25%	100%
6.5%	15.48%	16.22%	17.44%	18.18%	16.09%	16.58%	100%

As can be seen, there is no general tendency to choose a special weight combination. Consequently, none are more susceptible to cause misspecification in the model by choosing a complex interaction model when there is no interaction present.

Model choice

As the different models are not all nested the likelihood ratio test statistics are not appropriate for comparison. Therefore the AIC criterion was chosen for model selection. By applying AIC immediately it was found that only 54.75% to 63.84% of the models were found to have no interaction term. The variation is due to the different censoring percentages in the simulated data sets which were analysed separately. The lower recognition rates are found in the data with less than 50% censoring.
The multiplicative model is chosen by AIC in 10.035% to 14.82%, the transformed product model in 9.23% to 14.015% and the newly introduced term model in 16.24% to 16.925% of the models.
Obviously, the AIC criterion often favours an interaction model although the corresponding additional term does not contribute significant information with respect to its likelihood ratio test statistic. So long as the AIC criterion is in favour of the main effect model none of the likelihood ratio test statistics would contradict this result.
A better way to realise model selection is to change the order of the use of AIC and likelihood ratio procedures. Initially, significant interaction models are detected by likelihood ratio. If there are no significant models, or only one, the procedure stops, and the main effect model, or the only significant interaction model respectively, is chosen. In this way 96.765% to 97.935% of the models are already detected in the first step. Otherwise, if there is more than one significant interaction model these models will be compared by AIC.

The two presented methods lead to the same results, but the latter is more straightforward. The number of models chosen based on combined likelihood ratio and AIC procedure was detected for each censoring percentage separately. The resulting percentages of models chosen within the different censoring data sets are shown in table 5.

Table 5. Results of model choice after likelihood ratio test and AIC procedure

% censoring	***no interaction***	f_m	$f_{t.m}$	f_c
82.2%	91.39%	2.205%	2.125%	4.28%
71.4%	91.415%	2.39%	2.225%	3.97%
63.6%	90.985%	2.61%	2.21%	4.195%
52.9%	89.825%	2.94%	2.8%	4.435%
37.8%	88.46%	3.325%	3.56%	4.655%
24.5%	89%	3.3%	3.275%	4.425%
12.0%	89.97%	3.11%	2.92%	4%
6.5%	90.12%	2.93%	2.88%	4.07%

All of the upper bound results for interaction models in table 5 are found within the 37.8% censoring data, whereas the lower bounds are found for the two highest censoring rates. As shown in table 3 the chosen critical values for a maximal type I error of 5% lead to type I error rates $< 5\%$ in all of the data sets with censoring percentage different from 37.8%. Therefore, more interaction models are rejected during the likelihood ratio test procedure than in the data set which provided the critical value. Consequently, the lowest number of models without an interaction term is detected in the 37.8% censoring data and the highest number in high censoring data (71.4%).
Significance for an interaction term not only occurs in the model selected by AIC but also in competing models. The combinations of significant interaction terms have been analysed. The simultaneous occurrence of significance for the multiplicative and for the transformed multiplicative interaction model was detected much more often than for any of the other combinations. This indicates that the new term can be more easily distinguished from the other interaction models.

2.5 Misspecification errors

In the next step, data was simulated for models containing an interaction term. Covariables and observation times were calculated as in the preceding section. The maximum censoring time was set to 1000. The influence of the interaction term, denoted here as w, was varied as an integer between -6 and 6. The data was obtained from 1000 simulated data sets, each consisting of 1000 observations, computed for each parameter w separately.

Model with multiplicative interaction term

Initially, a multiplicative interaction term was included in the model. The corresponding linear predictor is:

$$lin.pred. = x_1 - 0.5 * x_2 + w * x_1 x_2$$

The resulting mean censoring percentage was between 36.33% and 43.89%. The analysis was carried out as in the preceding section using the earlier calculated adjusted critical values for testing significance of the different interaction terms. In case of more than one significant interaction term, the AIC criterion lead to the final model choice. The correct interaction term could be detected in 100% of the models. Although, the transformed multiplicative interaction term was also found significant in all cases, whereas a complex interaction term could be found significant in more than 99% of cases for $w \in \{1; 2; 3; 4\}$, in 95% to 97.6% of cases for $w \in \{-3; -2; -1; 5\}$ and in 75.6% to 90.7% of cases for $w \in \{-6; -5; -4; 6\}$. Consequently the choice is mainly based on AIC. If the simple multiplicative model would not be taken into account a different interaction surface would be chosen, i.e. an interaction effect would, at least, be detected.
For all positive values of w, the preferred weight combination within a significant complex interaction term was (1;-1). For negative values of w, except for $w = -6$ where again (1;-1) was preferred, the combination (2;0.5) was chosen more often than any other. This tendency, which is more obvious for increasing absolute values of w, is interesting although the final result of the created model choice algorithm is not affected.

Model with transformed multiplicative interaction term

In the second step, the underlying model for the data simulations contained a transformed multiplicative interaction term. Hence, the linear predictor is defined as:

$$lin.pred. = x_1 - 0.5 * x_2 + w * f_{logistic}(x_1) * f_{logistic}(x_2)$$

The resulting censoring percentage has a wider range, between 17.77% and 69.13%. This large range results because the uniformly distributed censoring time was kept fixed with a maximum of 1000 for all the simulations while the relative risk is varying due to the variation of w.
The results are summed together for the different values of w in table 6. The column *% AIC decisions* denotes the percentage of decisions using AIC i.e. the percentage of decisions which could not be made in the first step, after evaluating the three likelihood ratio tests. Thus, the number of AIC decisions is a measure for the ease with which the final model can be detected.

Table 6. Results of model choice based on LR and AIC for underlying interaction $f_{t.m}$

w(**% censoring**)	**% AIC decisions**	**model** $f_{t.m}$	**no int**	**model** f_m	**model** f_c
-6(69.13%)	94.9%	80.1%	1.3%	16.2%	2.4%
-5(64.87%)	89.8%	74.1%	5.2%	18.6%	2.1%
-4(60.03%)	82.1%	68.6%	8.3%	19.8%	3.3%
-3(54.63%)	58%	49.8%	25.6%	19.6%	4%
-2(48.96%)	34%	28.1%	51.8%	16.2%	3.9%
-1(43.32%)	10%	11.2%	76.7%	7.1%	5%
1(32.91%)	11.5%	11.8%	75.5%	8%	4.7%
2(28.78%)	45.1%	40.2%	39.2%	17%	3.6%
3(25.78%)	83.5%	72%	6.5%	18.8%	1.7%
4(22.36%)	98.2%	85.1%	0.5%	13.8%	0.6%
5(19.88%)	100%	90%	–	10%	–
6(17.77%)	100%	94.3%	–	5.7%	–

As can be seen, the correct interaction term is specified more easily when influence w is high. The number of AIC decisions increases with w, which is directly related to the decreasing number of models that do not show significance for any interaction term as these do not need a decision based on AIC. For low absolute values of w many models do not recognise any of the offered interaction terms.
As far as problematic weight combinations within significant complex interaction terms are concerned, a tendency to choose (2;-0.5) for positive values of w and (0.5;2) for negative values of w occurs, increasing with absolute values of w. The percentage of chosen complex interaction models, however, does not exceed 5% whereas the percentage of chosen simple variable product models f_m reaches up to 19.8%. Considering that the transformed variable product was found to be a significant term throughout the simulations based on an underlying variable product term in the previous section

and considering the remark at the end of the section 2.4 where both of these terms often showed significance at the same time although there was no interaction term included in the simulation, they seem to be much more similar to each other than to the complex interaction term.

Model with complex interaction term

Finally, data was simulated based on models containing the newly introduced complex interaction term, for all six of the weight combinations separately, resulting in the following linear predictor:

$$lin.pred. = x_1 - 0.5 * x_2 + w * f_{logistic}\left(w_1 x_1 + w_2 x_2\right)$$

The mean censoring percentage was between 4.79% and 30.52% for positive values of w and 48.53% to 92.14% for negative values of w.
As in the analysis above, all weight combinations were checked for influence $w \in \{-6; -5; -4; -3; -2; -1; 1; 2; 3; 4; 5; 6\}$. For all of the combinations of the six weights within the interaction term and influence parameter w 1000 data sets consisting of 1000 observations were simulated. The results for the different weight combinations are displayed in detail in table 7 (2;0.5), table 8 (0.5;2), table 9 (1;1), table 10 (1;-1), table 11 (-0.5;2) and table 12 (2;0.5) on the following pages. The percentage of completely correct model specifications, i.e. complex model with correct weight combination inside the interaction term, is indicated in brackets next to the percentage of decisions for a complex model. Again, the percentage of model choices which demanded a decision based on AIC is displayed in the column *% AIC decisions*.

Table 7. Results of model choice based on LR and AIC for underlying interaction $f_c = f_{logistic}(2x_1 + 0.5x_2)$. The number in brackets next to the percentage of identified complex interaction models indicates percentage of models with completely correct specification (i.e. including weight combination in the interaction term).

w**(% cens)**	**% AIC decisions**	f_c **(% (2;0.5))**	**no int**	f_m	$f_{t.m}$
-6(88.06%)	47.3%	99.6% (98.5%)	0.3%	–	0.1%
-5(84.89%)	37.5%	99.5% (98.1%)	0.4%	0.1%	–
-4(79.67%)	22%	98.6% (96%)	0.9%	0.1%	0.4%
-3(71.71%)	11.6%	94.7% (89.9%)	4.3%	0.3%	0.7%
-2(60.70%)	7.6%	78.8% (70.7%)	18.4%	1.7%	1%
-1(48.59%)	4.1%	27.2% (17.7%)	67.3%	3.7%	1.8%
1(29.91%)	8.3%	41.9% (30.7%)	51.3%	5.1%	1.7%
2(24.41%)	22.3%	94.7% (90.1%)	4.0%	1.2%	0.1%
3(20.92%)	40.9%	100% (99.5%)	–	–	–
4(18.34%)	59.1%	100% (100%)	–	–	–
5(16.43%)	71.0%	100% (100%)	–	–	–
6(14.88%)	78.4%	100% (100%)	–	–	–

Table 8. Results of model choice based on LR and AIC for underlying interaction $f_c = f_{logistic}(0.5x_1 + 2x_2)$. The number in brackets next to the percentage of identified complex interaction models indicates percentage of models with completely correct specification (i.e. including weight combination in the interaction term).

w**(% cens)**	**% AIC decisions**	f_c **(% (0.5;2))**	**no int**	f_m	$f_{t.m}$
-6(78.59%)	99.7%	98.6% (98.5%)	–	–	1.4%
-5(75.50%)	99.7%	98.8% (98.4%)	–	–	1.2%
-4(71.56%)	98%	97.3% (96%)	0.1%	0.1%	2.5%
-3(66.00%)	84.7%	92.7% (88.8%)	2.1%	0.2%	5%
-2(58.35%)	44.7%	78.7% (70.4%)	13%	0.9%	7.4%
-1(48.63%)	10.7%	32% (22.3%)	59.9%	1.7%	6.4%
1(27.85%)	7.1%	41.0% (30.1%)	54.9%	1.6%	2.5%
2(20.00%)	24.7%	98.5% (94.6%)	1.3%	–	0.2%
3(14.89%)	44.6%	100% (99.8%)	–	–	–
4(11.58%)	65.7%	100% (100%)	–	–	–
5(9.38%)	79.5%	100% (100%)	–	–	–
6(7.95%)	92.1%	100% (100%)	–	–	–

Table 9. Results of model choice based on LR and AIC for underlying interaction $f_c = f_{logistic}(x_1 + x_2)$. The number in brackets next to the percentage of identified complex interaction models indicates percentage of models with completely correct specification (i.e. including weight combination in the interaction term).

w(% cens)	% AIC decisions	f_c (% (1;1))	no int	f_m	$f_{t.m}$
-6(85.82%)	64.9%	80.3% (76.2%)	8.4%	3.1%	8.2%
-5(81.93%)	53.3%	76.3% (70.1%)	13.5%	2.9%	7.3%
-4(76.61%)	32.3%	68.9% (62.1%)	23.1%	2.5%	5.5%
-3(69.17%)	17.4%	54.1% (46.4%)	38.8%	1.9%	5.2%
-2(59.56%)	7.8%	29% (21.7%)	63.4%	3.3%	4.3%
-1(48.64%)	3%	12% (6.3%)	83.1%	2.4%	2.5%
1(28.81%)	4.6%	14.7% (8.2%)	78.6%	3.9%	2.8%
2(21.94%)	7.6%	53.9% (44.2%)	43.1%	2.2%	0.8%
3(17.03%)	26.8%	91.1% (84.2%)	7.4%	1.2%	0.3%
4(13.65%)	37.9%	99.3% (98.1%)	0.5%	0.1%	0.1%
5(11.16%)	50.4%	100% (99.9%)	–	–	–
6(9.38%)	64.3%	100% (99.9%)	–	–	–

Table 10. Results of model choice based on LR and AIC for underlying interaction $f_c = f_{logistic}(x_1 - x_2)$. The number in brackets next to the percentage of identified complex interaction models indicates percentage of models with completely correct specification (i.e. including weight combination in the interaction term).

w(% cens)	% AIC decisions	f_c (% (1;-1))	no int	f_m	$f_{t.m}$
-6(92.14%)	8%	54.1% (43.6%)	40.2%	3.3%	2.4%
-5(88.58%)	5.8%	53.4% (44.2%)	43.6%	1.8%	1.2%
-4(82.68%)	5.9%	56.2% (47.6%)	41.1%	1.5%	1.2%
-3(73.57%)	8.4%	50.4% (40.5%)	45.7%	1.8%	2%
-2(61.39%)	6.4%	29.3% (20.9%)	64.7%	4%	2%
-1(48.55%)	3.4%	10% (5.3%)	84.5%	3.1%	2.4%
1(29.97%)	5.1%	13.3% (7.3%)	79.9%	4.1%	2.7%
2(24.20%)	11.8%	42.4% (33.7%)	50.6%	4.9%	2.1%
3(20.21%)	14.9%	77.8% (72.3%)	18.9%	2.4%	0.9%
4(17.10%)	34.6%	94.9% (92%)	4.4%	0.4%	0.3%
5(14.85%)	46.2%	99.4% (99%)	0.3%	0.2%	0.1%
6(13.04%)	49.8%	99.5% (99%)	0.3%	0.1%	0.1%

Table 11. Results of model choice based on LR and AIC for underlying interaction $f_c = f_{logistic}(-0.5x_1 + 2x_2)$. The number in brackets next to the percentage of identified complex interaction models indicates percentage of models with completely correct specification (i.e. including weight combination in the interaction term).

w**(% cens)**	**% AIC decisions**	f_c **(% (-0.5;2))**	**no int**	f_m	$f_{t.m}$
-6(75.42%)	80.9%	100% (99.8%)	–	–	–
-5(72.64%)	71.7%	99.8% (99.7%)	–	–	0.2%
-4(69.12%)	60.1%	99.8% (98%)	0.1%	–	0.1%
-3(64.16%)	40.4%	98.3% (93.4%)	1.3%	–	0.4%
-2(57.52%)	20%	81.2% (72.6%)	15.7%	1%	2%
-1(48.66%)	5.1%	33.2% (23.3%)	62.2%	2%	2.6%
1(26.80%)	4.3%	41.2% (31.4%)	55.4%	0.9%	2.5%
2(17.63%)	23.4%	99.2% (96%)	0.4%	–	0.4%
3(11.61%)	58.3%	100% (100%)	–	–	–
4(8.07%)	85.8%	100% (100%)	–	–	–
5(6.05%)	97.3%	100% (100%)	–	–	–
6(4.79%)	99.2%	100% (100%)	–	–	–

Table 12. Results of model choice based on LR and AIC for underlying interaction $f_c = f_{logistic}(2x_1 - 0.5x_2)$. The number in brackets next to the percentage of identified complex interaction models indicates percentage of models with completely correct specification (i.e. including weight combination in the interaction term).

w**(% cens)**	**% AIC decisions**	f_c **(% (2;-0.5))**	**no int**	f_m	$f_{t.m}$
-6(90.84%)	13.9%	98.5% (95.2%)	1.3%	0.1%	0.1%
-5(87.73%)	12.4%	98.1% (93.8%)	1.7%	0.2%	–
-4(82.48%)	17.7%	96.3% (91.9%)	2.8%	0.3%	0.6%
-3(74.01%)	18.3%	95.5% (89.6%)	4%	–	0.5%
-2(61.85%)	13.2%	79.6% (71%)	17.8%	0.3%	2.3%
-1(48.53%)	5.5%	27.9% (19%)	66.3%	1.3%	4.5%
1(30.52%)	10.3%	35.2% (26.8%)	57.4%	2.2%	5.2%
2(25.49%)	24.4%	91.3% (87.4%)	7.3%	0.2%	1.2%
3(22.18%)	43.6%	99.8% (99.4%)	0.1%	–	0.1%
4(19.73%)	63.5%	100% (100%)	–	–	–
5(17.80%)	73.9%	100% (99.9%)	–	–	–
6(16.34%)	81.0%	100% (100%)	–	–	–

The low number of AIC based decisions in combination with high numbers of models chosen to have no interaction term in tables 7-12 especially found for low influence $|w|$ indicate that complex interaction terms are difficult to detect. The risk of specification of a model containing one of the two alternative interaction terms is rather low compared to the risk of not detecting any interaction term. Most model choices are already complete after evaluating the different likelihood ratio test statistics. In most of the cases a complex interaction term is the only interaction term which is found significant or a main effect model is preferred. The limit of 5% misspecification rate for each of the alternative interaction terms is rarely exceeded.
In no more than 11.2% of cases in any simulation run, a complex interaction model with wrong weight combination is chosen. This percentage includes all wrong weight combinations which occurred during the corresponding simulation and has therefore been studied more closely. Four times the 5% limit of chosen models was slightly exceeded for a single weight combination different from that underlying the simulation. The maximum misspecification error for a single weight combination was 5.8%, which is still low. It occurred for underlying interaction term $-6 * f_{logistic}(x_1 - x_2)$. The determined ideal combination was (-0.5;2).
The maximum sum of wrong interaction terms detected as best fitting per simulation set, i.e. per line in tables 7-12, including simple product, transformed product and complex interaction term with wrong weight combination, for each underlying weight combination ranged between 14.5% and 18.0%. This maximum misspecification rate was found for most weight combinations within the simulation studies for $w \in \{-1; 1\}$. For combinations (1;1) and (1;-1) this maximum of misspecification error was reached for $w = -5$ and for $w = -6$ respectively. For most w, except for high positive values, for the latter combinations the misspecification rate was more than 10% in contrast to the other combinations for which the 10% are not exceeded more than three times.
The percentage of completely correct model specifications are displayed in figure 5 for models containing interaction terms either of the form $f_{t.m}$ or f_c, the latter for all predefined weight combinations. For low values of w it can therefore be stated that correct model determination is difficult for these interaction surfaces. For interactions f_m, all models could be specified correctly for all influence parameters w. Therefore, these results have not been included in figure 5.

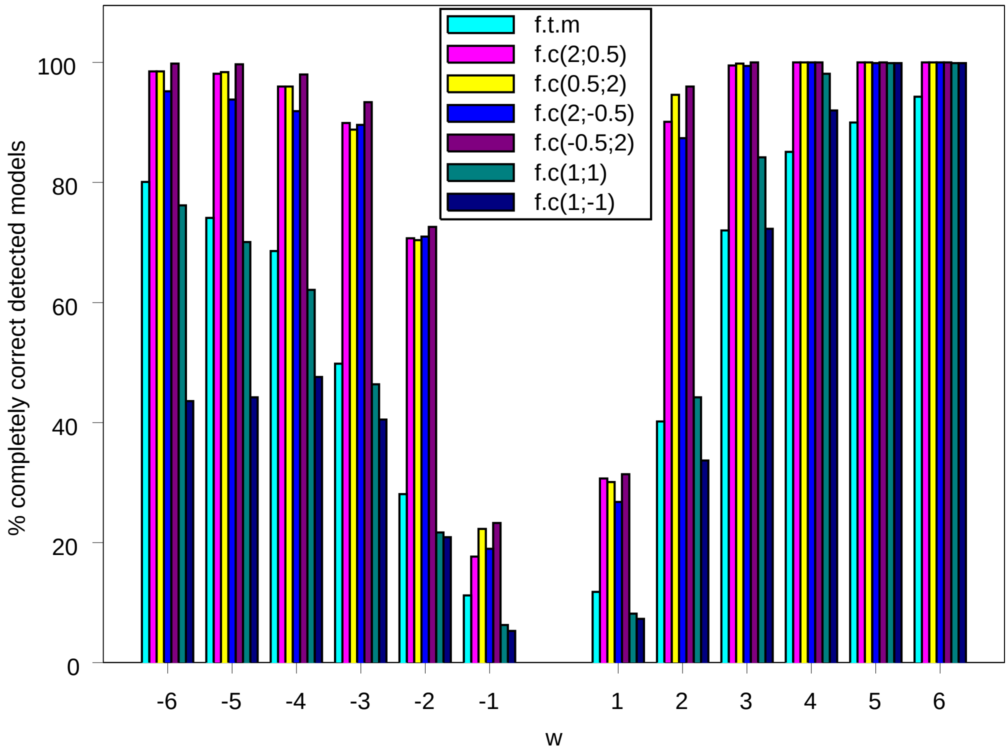

Figure 5. Percentages of true model specifications for simulated survival data containing interaction terms of the form $f_{t.m}$ or f_c (modelled for the six inner weight combinations separately) in the linear predictor. The results are displayed clustered for different interaction influence parameters w. Model choice is based on likelihood ratio tests and AIC criterion.

3 Conclusion

The Cox model has been extended by some new interaction terms. The initial goal was to implement the concept of a logistic activation function on weighted inputs of neural networks into standard methods for survival analysis. Hence, the new interaction terms include logistic transformations and are defined as: $f_{t.m} = f_{logistic}(x_1) * f_{logistic}(x_2)$ and $f_c = f_{logistic}(w_1 * x_1 + w_2 * x_2)$. For the latter a predefined set of weights w_1 and w_2 was found.
While checking the critical values of the newly introduced terms as well as that of the standard product interaction term, $f_m = x_1 * x_2$, by likelihood ratio tests allowing for only 5% type I error a deviation from the expected χ^2 distributions was observed. The desired critical values seem to depend on the censoring rate in the analysed data set. The exact relationship is subject of further research. New critical values have been defined for the likelihood ratio test statistics as the maximum 95% quantile of which in

eight simulation studies with different mean censoring rates.
The new critical values in combination with the AIC criterion were then used to define a model choice algorithm which was applied in different survival data simulations for different underlying models. Each of the latter contained one of the new interaction terms or the standard variable product.
It was found that the new interaction terms can be detected better if they are of high influence. Otherwise, in most cases, a model without any interaction would be preferred. The misspecification rate, i.e. detection of a wrong interaction model, is also increasing for low influence of the interaction term. The terms $f_{t.m}$ and f_m seem to be more related to each other than to f_c which results in high numbers of simultaneous significances for the terms in different settings. This also means that an interaction effect can often be detected by checking only for f_m, which is a standard method in medical research, even if the true interaction has the form $f_{t.m}$. Therefore, interpretation can be misleading. From f_c, however, both of these interaction terms can be distinguished more clearly. Thus, for f_c, the danger of choosing a weight combination inside the term that is different to that defined for the underlying term is often higher than for choosing a completely different interaction model.
Consequently, the newly introduced interaction surfaces can help to detect underlying interaction structures and prevent misleading results caused by only checking for a standard variable product term. Additionally, a model choice algorithm for detection of the best interaction surface can easily be implemented into standard software and made available for medical researchers. A subject of further research is the determination of more interaction surfaces resulting from polynomial transformations of the corresponding variables. The observed relationship between likelihood ratio test statistic and censoring percentage is to be studied more in detail.

References

[1] U. Anders (1997). Statistische neuronale Netze. Verlag Vahlen. München

[2] U. Berger, J. Schäfer, K. Ulm (2002). Dynamic Cox modelling based on fractional polynomials: time-variations in gastric cancer prognosis. *Stat. in Med.*, 22(7), 1163-1180

[3] N.E. Breslow (1974). Covariance analysis of censored survival data. *Biometrics*, 30, 89-100

[4] D.R. Cox (1972). Regression models and life-tables (with discussion). *J. Royal Stat. Soc. B*, 34:187-220

[5] W. Duch and N. Jankowski (1997). New neural transfer functions. *Journal of Applied Mathematics and Computer Science*, 7, 639-658

[6] L. Fahrmeir, A. Hamerle, G. Tutz (1996). Multivariate statistische Verfahren. Walter de Gruyter & Co. Berlin, New York

[7] P.M. Grambsch, T.M. Therneau (1995). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81, 515-526

[8] T. Hastie, R. Tibshirani (1993). Varying-coefficient models. *J. Royal Stat. Soc. B*,55, 757-796

[9] Lang, S. and Brezger, A. (2003). Bayesian P-Splines. *Journal of Computational and Graphical Statistics*, to appear

[10] M. Leblanc, J. Crowley (1999). Adaptive regression splines in the Cox model. *Biometrics*, 55, 204-213

[11] R. Maller, X. Zhou (1996). Survival analysis with long-term survivors. John Whiley & Sons Ltd, Chichester

[12] B.D. Ripley (1996). Pattern recognition and neural networks. Cambridge University Press. Cambridge

[13] R.M. Ripley (1998). Neural network models for breast cancer prognosis. *Dissertation*. University of Oxford

[14] P. Royston (2002). The use of fractional polynomials to model interactions between treatment and continuous covariates in clinical trials. Stata User's Group, 21 May 2002

[15] G. Salanti (2003). The isotonic regression framework - Estimating and testing under order restrictions. *Dissertation*, Ludwig-Maximilians-Universität, München

[16] G. Salanti, K. Ulm (2003) Modelling time-varying effects in Cox model under order restrictions. Discussion paper 234, Sonderforschungsbereich 386, Ludwig-Maximilians-Universität München

[17] G. Schwarzer, W. Vach, M. Schumacher (2000). On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Statistics in Medicine*, 19, 542-561

[18] T. Therneau, P. Grambsch (2000). Modelling Survival Data - Extending the Cox Model. Springer-Verlag New York Berlin Heidelberg

[19] T. Therneau, P. Grambsch, T. Fleming (1990). Martingale based residuals for survival models. *Biometrika*, 77, 147-160

[20] R. Xu, S. Adak (2002). Survival analysis with time-varying regression effects using a tree-based approach. *Biometrics*, 58, 305-315

[21] H. Zhang, B. Singer (1999). Recursive Partitioning in Health Sciences. Springer-Verlag, New York, Inc.